Long-Time Protein Folding Dynamics from Short-Time Molecular Dynamics Simulations

ARTICLE in SIAM JOURNAL ON MULTISCALE MODELING AND SIMULATION - JANUARY 2006
Impact Factor: 1.63 · DOI: 10.1137/06065146X

CITATIONS 115  READS 36

4 AUTHORS, INCLUDING:

John Damon Chodera
Memorial Sloan-Kettering Cancer Center
64 PUBLICATIONS 3,672 CITATIONS

William C Swope
IBM
77 PUBLICATIONS 6,193 CITATIONS

Ken A Dill
Stony Brook University
392 PUBLICATIONS 28,354 CITATIONS

All in-text references underlined in blue are linked to publications on ResearchGate, letting you access and read them immediately.

Available from: Ken A Dill
LONG-TIME PROTEIN FOLDING DYNAMICS FROM SHORT-TIME
MOLECULAR DYNAMICS SIMULATIONS

JOHN D. CHODERA†, WILLIAM C. SWOPE‡, JED W. PITERA‡, AND KEN A. DILL§

Abstract. Protein folding involves physical timescales—microseconds to seconds—that are too
long to be studied directly by straightforward molecular dynamics simulation, where the fundamental
timestep is constrained to femtoseconds. Here we show how the long-time statistical dynamics of a
simple solvated biomolecular system can be well described by a discrete-state Markov chain model
constructed from trajectories that are an order of magnitude shorter than the longest relaxation
times of the system. This suggests that such models, appropriately constructed from short molecular
dynamics simulations, may have utility in the study of long-time conformational dynamics.

Key words. Markov chain model, molecular dynamics, peptide dynamics, protein folding

AMS subject classifications. 60J10, 60J20, 80A30, 92C05, 92C45, 70K70, 74A25

DOI. 10.1137/06065146X

1. Introduction. Proteins can fold to well-defined native topologies with surprising
determinism. Many small, single domain proteins can fold rapidly, reversibly,
cooperatively, and without the aid of other molecular machinery. In response to an
environmental perturbation such as the introduction or removal of denaturant or a rapid
change in solvent temperature, these fast-folding proteins exhibit nearly exponential
relaxation kinetics with observed time constants on the order of microseconds. Other
proteins exhibit slow and complex kinetics, suggesting the presence of one or more
kinetic intermediates. A detailed understanding of this process has been the focus of
much of modern biophysics. Ultimately, knowledge of the general mechanical features
by which proteins fold and aggregate is critical for understanding a variety of folding
and misfolding diseases, elucidating principles necessary for effective protein design,
and developing the basic tools needed for other related technological applications of
complex molecular structures.

A description of the mechanism by which a particular protein folds must by
necessity be a statistical one. While the initial microscopic state and dynamical
trajectory may differ for each molecule in an experiment, many proteins refold to
their native (folded) topologies upon the restoration of native conditions with the
certainty of macroscopic law. A proper statistical description would summarize
the salient features and relative probabilities of relevant folding routes in a way that

*Received by the editors February 2, 2006; accepted for publication (in revised form) May 5, 2006; published electronically December 28, 2006.
†Corresponding author. Graduate Group in Biophysics, University of California at San Francisco, 600 16th St., Box 2240, San Francisco, CA 94143-2240 (jchodera@gmail.com). This author was
supported by Howard Hughes Medical Institute and IBM predoctoral fellowships.
‡IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120 (swope@almaden.ibm.com,
pitera@us.ibm.com). The second author was supported in part by National Science Foundation
MRSEC Center on Polymer Interfaces and Macromolecular Assemblies grant DMR-0213618.
§Department of Pharmaceutical Chemistry, University of California at San Francisco, 600 16th St.,
Box 2240, San Francisco, CA 94143-2240 (dill@zimm.compbio.ucsf.edu). This author was supported
by National Institutes of Health grant GM34993.

1By native topology or native structure, we refer to the ensemble of configurations sharing a
coarse overall structure, or fold, with the experimental structure.

2By microscopic state, we refer to the set of generalized coordinates and momenta that completely
determine the microscopic state of the system, such as the phase space point.
is meaningful to the physical chemist. This manner of model has been difficult to extract from experiments. Despite the high time resolution possible with optical spectroscopy, the majority of these experiments rely on the observation of an ensemble of molecules to obtain sufficient signal, resulting in the ability to observe only (possibly time-dependent) ensemble averages, rather than the behavior of any single molecule. While observations of single molecules are now possible with fluorescence techniques, atomic force microscopy, or optical traps, high temporal resolution is sacrificed to achieve sufficient signal for reliable measurement. In contrast, computer simulation promises the ability to produce information with both atomic detail and high temporal resolution.

In practice, however, the presence of fast vibrational motion constrains the fundamental integration timestep to femtoseconds in order to ensure stability, limiting practical straightforward molecular dynamics simulations of atomically detailed representations of solvated proteins to tens of nanoseconds. As even the fastest folding proteins exhibit relaxation timescales of several microseconds [24], this leads to a timescale gap of at least two decades in time. Using supercomputers such as Blue Gene [18] and software specialized for molecular dynamics simulations on these computer systems [17, 19], one can produce atomistic simulations of protein molecules with explicit representation of surrounding solvent on several microsecond timescales. However, the number of trajectories that need be generated to provide an adequate statistical characterization of the folding mechanism of even a single protein makes such an endeavor extremely challenging. Distributed computing projects such as Folding@Home [36] regularly collect tens of thousands of trajectories tens of nanoseconds in length, but extracting insight about microsecond timescale dynamics from these large datasets can be difficult [16, 35, 30].

Kinetics models may provide the necessary link between short simulations of a single molecule and long experimental observations of ensembles of molecules. If time evolution of a protein system is characterized by long waiting times within metastable states punctuated by infrequent transitions between these states, interstate dynamics may appear stochastic and memoryless on some short timescale. In this case, long trajectories may be modeled as a Markov chain realized on a discrete state space of a (hopefully small) number of states. While this model could not describe dynamical behavior at very short timescales, which is dominated by molecular motion within a metastable state, it could nevertheless faithfully describe long-timescale transitions between states. This approach would have numerous advantages. It is precisely these slow transitions involving major structural rearrangements that are of primary interest; elimination of high-frequency detail is often desirable in aiding interpretation of trajectories. To generate a statistical description of folding dynamics, instead of generating many simulations, each long enough to contain complete folding events, we need only generate simulations long enough to characterize transition rates between pairs of conformational substates. Construction would therefore be amenable to parallelization on loosely coupled grids of computer systems. The resulting kinetic model could then be used to compute the stochastic temporal evolution of either a single molecule or an ensemble of molecules, allowing direct comparison to data from both kinds of kinetics experiments, or to answer statistical questions about folding pathways and mechanisms that are currently experimentally inaccessible.

This proposition is not entirely novel. Several groups have constructed stochastic kinetic models from states defined by local potential energy minima of small peptides, using transition state theory to estimate interstate transition rates [9, 26, 3, 28, 33, 32]. Unfortunately, the number of minima grows rapidly with increasing system size, mak-
ing the procedure prohibitively expensive for larger proteins or systems containing explicit solvent molecules. Other work [11, 50, 46, 1, 47, 40] has focused on the construction of discrete- or continuous-time Markovian models to describe dynamics between a small number of states. These models, however, have yet to demonstrate that they can adequately describe the dynamics on timescales much longer than the trajectories from which the models were constructed; no attempt is made to compare the dynamics predicted by the model with long trajectories of explicitly solvated systems. Transition interface sampling [31] and “milestoning” [15] attempt to describe dynamics along a one-dimensional reaction coordinate, but these approaches are valid only if it can be shown that relaxation transverse to this coordinate happens very quickly. Some have suggested constructing stochastic models of dynamics by expansion of the dynamical operator in a smooth basis set [43, 51, 44], but this approach unfortunately suffers from the great difficulty of choosing rapidly convergent basis sets for large molecules. In spite of the challenges with their construction, Markov models that can accurately capture the long-time kinetics of a system can be very useful. Such models embody a concise description of the various kinetic pathways and their relative likelihood. Moreover, they can be used to facilitate the computation of useful properties such as state lifetimes [49], mean first-passage times [46], the existence of hidden intermediates [34], and $\mathcal{P}$ fold values (transmission coefficients) [27].

Here we present a proof of principle for how the dynamics of a solvated biomolecular system can be described using information from short simulations. This is illustrated using terminally blocked alanine, a system small enough that its dynamical behavior can also be thoroughly characterized by straightforward molecular dynamics simulation. First, a parallel tempering simulation is conducted to explore the thermally relevant regions of configuration space. Next, a set of metastable states, corresponding to regions of configuration space with low probabilities of escape, are identified. Due to the simplicity of the system considered in this work, these states can readily be identified by hand. This removes the complication of choice of state decomposition for arbitrary macromolecular systems, which we shall not address here. Finally, a number of short trajectories are initiated from each state, and a Markov chain model is constructed from analysis of the observed interstate transitions. We demonstrate the validity of the model by comparing its prediction for the long-time evolution of a nonequilibrium ensemble with what is actually exhibited by an ensemble of long simulations.

This paper is organized as follows. In section 2, the Markov chain model and its method of construction are described. In section 3, the method is applied to terminally blocked alanine in explicit solvent. A discussion of the significance of this result, as well as problems remaining to be solved before the method can be applied to larger biomolecules, follows in section 4.

2. Theory.

2.1. Conformational dynamics as a Markov process. Consider the dynamics of a macromolecular system in equilibrium at some temperature of interest, where we have decomposed all configuration space into a set of $M$ disjoint but contiguous states. If we observe a trajectory of this system at times $t = 0, \tau, 2\tau, \ldots, n\tau$, where $\tau$ denotes the observation interval, we can represent the trajectory in terms of the state the system occupies at each of these discrete times, $s_0, s_1, s_2, \ldots, s_n$. The sequence of states produced by such a trajectory is a discrete-time stochastic process. If this process is a Markov chain, it must satisfy the Markov property, whereby the probability of observing the system in state $s_n$ at the time point $n\tau$, given the state history
$s_0, s_1, s_2, \ldots, s_{n-1}$, is independent of all but the previous state $s_{n-1}$. For a stationary process which has no explicit dependence on time, this property is given by

$$P(s_n|s_0, s_1, s_2, \ldots, s_{n-1}) = P(s_n|s_{n-1}).$$

As there are a finite number of states, this process can be entirely characterized by an $M \times M$ transition matrix $T(\tau)$ dependent only on lag time $\tau$. The element $T_{ji}(\tau)$ denotes the probability of observing the system in state $j$ at time $\tau$, given that it was initially in state $i$ at time 0:

$$T_{ji}(\tau) \equiv P(j|i).$$

If we do not know the precise initial state of the system at time 0 but only the probability the system started in each state, or if we observe an ensemble of many noninteracting systems in an experiment, we can instead consider the probability of finding one particular molecule in each state $i$ at time $n\tau$ as components of the vector of state probabilities $p(n\tau)$. If the initial probability vector is given by $p(0)$, we can write the probability vector at some later time $n\tau$ as

$$p(n\tau) = T(n\tau)p(0) = [T(\tau)]^n p(0).$$

This property is described by the Chapman–Kolmogorov equation.

### 2.2. Construction of the Markov chain model from simulation.

For a system in which the dynamical evolution is Newtonian but the initial configurations come from a canonical distribution, Swope, Pitera, and Suits [49] show that the transition probability $T_{ji}(\tau)$ can be written as

$$T_{ji}(\tau) \equiv \langle \chi_j(z(\tau)) \chi_i(z(0)) \rangle / \langle \chi_i(z(0)) \rangle = \frac{\int dz(0) e^{-\beta H(z(0))} \chi_j(z(\tau)) \chi_i(z(0))}{\int dz(0) e^{-\beta H(z(0))} \chi_i(z(0))} = \int dz(0) p_i(z(0)) \chi_j(z(\tau)),$$

where $z$ denotes a point in phase space, $\chi_i(z)$ denotes the indicator function for state $i$, $\beta \equiv (k_B T)^{-1}$ is the inverse temperature, $H(z)$ is the Hamiltonian, and $p_i(z)$ denotes the canonical distribution restricted to state $i$:

$$p_i(z) = \frac{e^{-\beta H(z)}}{\int dz e^{-\beta H(z)} \chi_i(z)}.$$

The final expression of (2.4) simply states the obvious: the transition matrix element $T_{ji}(\tau)$ can be estimated in a straightforward (though potentially inefficient) manner by initiating a number of simulations from configurations selected from a canonical distribution within state $i$, evolving the dynamics for a time $\tau$, and determining the fraction of trajectories that terminate in state $j$:

$$T_{ji}(\tau) \approx \frac{N_{ji}(\tau)}{\sum_{j'=1}^M N_{jj'}(\tau)}.$$

Here $N_{ji}(\tau)$ denotes the number of trajectories initiated from state $i$ that terminate in state $j$ at time $\tau$. This procedure corresponds to the method proposed earlier by

Downloaded 03/30/15 to 129.49.83.141 Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php
We do not expect dynamics of a macromolecule in solution to resemble a Markov process for all observation intervals \( \tau \), as ballistic motion dominates on very short times, and sufficient time must be allowed for collisions with the solvent and decorrelation of the trajectory within a metastable state. Imperfect definitions of the metastable states may also lead to non-Markovian behavior on short times [49]. At sufficiently long intervals \( \tau \), however, we might observe that dynamics resembles a Markov process. While it is impractical to test the condition of complete history independence (see (2.1)), we can simply check the (weaker) condition imposed by the Chapman–Kolmogorov equation (see (2.3)): For transition matrices constructed for a given \( \tau \), we check whether (2.3) holds for several lag times \( n = 2, 3, 4, \ldots \) to within statistical uncertainty. If so, the Markovian model can be assumed to be a reasonable model of dynamics.

3. Application to terminally blocked alanine peptide.

3.1. System setup and equilibration. Using the LEaP program from the AMBER7 molecular mechanics package [6], a terminally blocked alanine peptide (sequence ACE-ALA-NME; see Figure 1) was generated in the extended conformation with peptide force field parameters taken from the AMBER parm96 parameter set [23]. The system was subsequently solvated with 431 TIP3P water molecules [21] in a truncated octahedral simulation box with dimensions chosen to ensure all box boundaries were at least 7 Å from any atom of the extended peptide. All minimization and
molecular dynamics simulations were conducted using the **sander** program from the AMBER7 package. Default nonbonded cutoffs were used, bonds to hydrogen were constrained with SHAKE using a tolerance of $10^{-8}$ [39], and long-range electrostatics were treated by the particle-mesh Ewald method [10] with the default settings.

The system was first subjected to 50 steps of steepest descent energy minimization, followed by 1000 steps of conjugate gradient optimization. To equilibrate the explicit solvent system to the appropriate volume, a 100 ps molecular dynamics simulation was performed with the temperature adjusted to 300 K and the pressure to 1 atm by the Berendsen weak-coupling algorithm [4] with temperature and pressure relaxation time constants of 1 ps and 0.2 ps, respectively. The simulation box was fixed at the final size obtained from this equilibration step, with a volume of $13232 \text{Å}^3$, in all subsequent simulations.

### 3.2. Parallel tempering

In order to broadly explore the configuration space of the peptide and ensure that all important conformational substates were located, a parallel tempering (or replica exchange among temperatures) molecular dynamics simulation [48] was conducted using a parallel **Perl** wrapper for the **sander** program [7]. Forty replicas were used, with replica temperatures exponentially distributed over the range 273–600 K, yielding an average exchange acceptance probability of about 50%. All momenta were reassigned from the Maxwell–Boltzmann distribution at the appropriate replica temperature after each exchange attempt, and constant-energy, constant-volume molecular dynamics with a 2 fs timestep was performed between exchange attempts. The algorithm used to select pairs of replicas for temperature exchange attempts starts from the highest temperature replica, attempts to swap the configuration for the next-lowest temperature replica using the Metropolis-like criteria, and proceeds down the temperatures in this manner. On the next iteration, swapping attempts start from the lowest temperature and proceed upward, and this alternation in direction is continued in subsequent pairs of iterations.

Starting all replicas from the volume-equilibrated configuration described above, 100 iterations were conducted with 1 ps between exchange attempts to equilibrate the replicas to their respective temperatures. This equilibration run was followed by a production run of 500 iterations with 20 ps between exchange attempts, a total of 10 ns/replica. Solute configurations and potential energies from the production run were written to disk every 0.1 ps, while full-system restart files were recorded every 1 ps for the purpose of starting new simulations from these configurations, as described in section 3.4.

### 3.3. State decomposition

The slow degrees of freedom for terminally blocked alanine peptide (neglecting those involving solvent motion) can be captured by the two backbone torsion angles labeled $\phi$ and $\psi$ (see Figure 1) [5, 29]. To this end, the potential of mean force at 302 K was computed from the parallel tempering data using the weighted histogram analysis method (WHAM) [25, 8] and is shown in Figure 1. Six free energy basins are readily visible, and rectangular regions around these basins were chosen for the decomposition of all of configuration space into a set of six states. State definitions are listed in Table 1 and plotted as thick dividing lines in Figure 1.

### 3.4. Construction of Markov chain model from short trajectories

To construct a Markov chain model of dynamics once the states were identified, the interstate transition probabilities were computed using the procedure described in section 2.2. A set of 1000 energy-conserving trajectories 10 ps in length were generated from a canonical distribution of initial conditions within each state. This initial
Table 1

State definitions for the manual decomposition of (φ, ψ)-space into metastable states and populations at 302 K.

<table>
<thead>
<tr>
<th>State</th>
<th>Label</th>
<th>φ,ψ</th>
<th>P&lt;sub&gt;eq&lt;/sub&gt;</th>
<th>P&lt;sub&gt;eq&lt;/sub&gt;[K]</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>C&lt;sub&gt;5&lt;/sub&gt;</td>
<td>[117, -105)</td>
<td>[28, -124)</td>
<td>.4787 (.0613)</td>
</tr>
<tr>
<td>2</td>
<td>P&lt;sub&gt;H&lt;/sub&gt;</td>
<td>[-105, 0)</td>
<td>[28, -124)</td>
<td>.4159 (.0486)</td>
</tr>
<tr>
<td>3</td>
<td>α&lt;sub&gt;P&lt;/sub&gt;</td>
<td>[117, -105)</td>
<td>[-124, 28)</td>
<td>.0425 (.0038)</td>
</tr>
<tr>
<td>4</td>
<td>α&lt;sub&gt;R&lt;/sub&gt;</td>
<td>[-105, 0)</td>
<td>[-124, 28)</td>
<td>.0588 (.0079)</td>
</tr>
<tr>
<td>5</td>
<td>C&lt;sub&gt;ax&lt;/sub&gt;</td>
<td>[0, 117)</td>
<td>[111, -5)</td>
<td>.0030 (.0015)</td>
</tr>
<tr>
<td>6</td>
<td>α&lt;sub&gt;L&lt;/sub&gt;</td>
<td>[0, 117)</td>
<td>[-5, 111)</td>
<td>.0011 (.0004)</td>
</tr>
</tbody>
</table>

*Corresponding state labels from [38]. Equilibrium probabilities at 302 K estimated from the replica exchange simulation by WHAM with corresponding uncertainties representing one standard deviation shown in parenthesis.

distribution was generated by selecting initial configurations from all replicas of the replica exchange simulation with a probability proportional to their weight used in computing canonical averages at 302K, as determined by WHAM, and assigning initial momenta from the Maxwell–Boltzmann distribution. For each lag time τ, an estimate of the transition probability T<sub>ij</sub>(τ) was obtained using (2.6). A bootstrap procedure [14], in which 200 replicates of 1000 trajectories from each state were chosen with replacement from the set of trajectories emanating from each state, was used to estimate the uncertainty in the observed transition probabilities.

The observed transition probabilities out of each state as a function of τ are shown in Figure 2, along with the corresponding equilibrium probabilities of each state determined from the replica-exchange simulation. None of the state populations reach their equilibrium values within 10 ps, indicating the slowest relaxation timescales are much longer, perhaps substantially so for trajectories originating from states 5 and 6.

Transition matrices at several lag times—0.1 ps, 1 ps, 6 ps, and 10 ps—are shown in Table 2.

3.5. Comparison with long trajectories. To determine the accuracy with which transition matrices constructed from different lag times from short (10 ps) simulations are able to reproduce the statistical dynamics over long times (approximately 100 ps), state populations for an ensemble of trajectories emanating from each state were computed from the model and compared to the observed time evolution of a separate set of long trajectories. For this comparison, 1000 trajectories 100 ps in length were initiated from each state, using the same protocol in section 3.4. Figure 3 shows the time evolution of state populations from these trajectories (along with corresponding uncertainties) as a function of time. Superimposed are state populations computed by (2.3) from the transition matrices constructed for different lag times τ from the short simulations described in section 3.4. These are connected by straight line segments solely to guide the eye; the model cannot make predictions for the populations at times that are not integral multiples of the lag time τ.

The transition probabilities are poorly reproduced in the model constructed with a lag time of only 0.1 ps. Apparently, this time is so short that the system does not behave in a Markovian manner on this timescale. At a lag time of 1 ps, the agreement between the model and long simulations is clearly better, though there are still visible systematic deviations. By a lag time of 6 ps, the agreement is excellent. The model constructed from a lag time of 10 ps also shows excellent agreement, but by this time, the temporal resolution has started to become rather poor. Information about the
system is known only for times that are integral multiples of 10 ps. One can imagine that the most useful model would be constructed from the shortest lag time at which dynamics is Markovian, as this model has the highest temporal resolution while still correctly describing long-time dynamics.

### 3.6. Long-time dynamics from the Markov chain model

As an illustration of the utility of the Markov chain model, Figure 4 depicts an artificial trajectory generated by realization of the Markov process, 10 ns in length, three orders of magnitude longer than the short trajectories used to construct the model. While statistical properties of the dynamics can also be extracted in other ways, such as through an eigenvalue decomposition, it may be useful to generate artificial trajectories and analyze them directly. Note the infrequent sampling of states 5 and 6, states with very small equilibrium probabilities, and the long dwell times in the region formed by stable states 1 and 2.

### 4. Discussion

We have demonstrated that a Markov model constructed from simulations roughly one order of magnitude shorter than the slowest relaxation time in the system is sufficient to capture the long-time dynamics of a simple biomolecular system, terminally blocked alanine peptide in explicit solvent. Instead of generating large numbers of long trajectories to statistically characterize dynamics, we require
only a sufficient number of trajectories to estimate transition probabilities between pairs of states. In addition, these trajectories need only be long enough for interstate dynamics to appear Markovian. Once so constructed, the model can be used to answer various questions of interest regarding the long-time statistical dynamics without the need to perform additional simulations.

While it is impossible to predict what the minimum trajectory length required for Markovian behavior will be for other, larger systems, it is important to recall that most proteins fold on the millisecond to second timescale. Even fast folding proteins can require tens of microseconds to fold [24]. To bring the treatment of these systems within the realm of feasibility, the Markov time would need to remain sufficiently short to allow the collection of a significant number of trajectories despite the presence of relaxation times many orders of magnitude longer. No statement can yet be made about the number of states necessary to model more complex systems or whether this number might make this approach prohibitively expensive.

The question of how best to validate a Markov chain model constructed from short trajectories without additional long-time information is a topic of active research. To determine the lag time to construct the transition matrix so that the Markov chain is an accurate description of long-time dynamics, it was necessary to compare to an additional set of long trajectories. This, of course, defeats the utility of a model constructed from short trajectories. Other methods, such as tests of eigenvalue behavior [49] or direct tests of Markovity [37], may provide alternatives.

In this work, we have avoided the issue of how best to define the number and location of states used for construction of the Markov model. Ideally, these states

---

### Table 2

Transition matrices at several lag times estimated from a set of 10 ps trajectories.

<table>
<thead>
<tr>
<th></th>
<th>0.1 ps</th>
<th>1 ps</th>
<th>6 ps</th>
<th>10 ps</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>T(0,1 ps)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.967 .041 .029 .002</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.030 .959 .003 .001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.003 .912 .022 .059</td>
<td>.975 .093  .015</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.030 .959 .003 .001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.003 .912 .022 .059</td>
<td>.975 .093  .015</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.003 .912 .022 .059</td>
<td>.975 .093  .015</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>T(1 ps)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.856 .161 .096 .011 .002 .004</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.130 .835 .008 .007 .001 .086</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.014 .002 .002 .004 .001 .017</td>
<td>.873 .966  .047</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.014 .002 .002 .004 .001 .017</td>
<td>.873 .966  .047</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>T(6 ps)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.642 .400 .190 .068 .010 .069</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.324 .586 .069 .043 .011 .268</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.023 .009 .373 .251 .017 .002</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.011 .005 .367 .637 .075 .004</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.011 .005 .367 .637 .075 .004</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>T(10 ps)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.573 .459 .232 .157 .022 .138</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.385 .520 .110 .072 .033 .333</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.018 .013 .286 .235 .030 .005</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.022 .008 .371 .535 .111 .009</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.001 .001 .745 .127 .005 .009</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>.002 .001 .745 .127 .005 .009</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*a* Blank entries denote estimated transition probabilities of zero.
will be significantly metastable so that the system rapidly loses memory of its previous location after entering a state, before making a transition to another state. For the system considered, the slow degrees of freedom were known beforehand, so the potential of mean force in these coordinates revealed a useful set of states. In more complex systems, the coordinates in which dynamics is slow will be much more difficult to discern; some automatic method for the identification of metastable states is necessary. Pure conformational clustering methods [22, 20] may prove to be inadequate because they neglect the true locations of kinetic barriers, but attempts to also consider kinetic relationships give promising results but have not yet been applied to large explicitly solvated systems [41, 42, 40]. This problem is the subject of work soon to be reported [45].

Here we employed the most straightforward approach to estimating interstate transition probabilities, whereby a large number of short trajectories are initiated
from equilibrium within each state. While this approach is amenable to distributed or grid computing, the metastable nature of well-chosen states will result in many of these trajectories simply remaining in their state of origin, rather than contributing to estimates of the off-diagonal elements of the transition matrix. It is precisely these off-diagonal elements that are critical in determining which trajectories through state space are most likely. Algorithms employing importance sampling techniques in trajectory space—such as transition path sampling [12], transition interface sampling [52], and the string method [13]—may provide an efficient way to compute these interstate transition probabilities.

Acknowledgments. JDC gratefully acknowledges Libusha Kelly (UCSF) for critical reading of this manuscript and Nina Singhal (Stanford) for stimulating discussion and insightful criticism.

REFERENCES

LONG-TIME DYNAMICS FROM SHORT-TIME SIMULATIONS


